Laser Printer Attribution: Exploring New Features and Beyond

Anselmo Ferreira¹, Luiz Navarro¹, Giuliano Pinheiro¹, Jefersson Dos Santos² and Anderson Rocha¹

¹Institute of Computing, University of Campinas
Av. Albert Einstein, 1251, Cidade Universitaria, Campinas/SP - Brasil, CEP 13083-852

²Universidade Federal de Minas Gerais, Department of Computer Science
Av. Antônio Carlos 6627 - Prédio do ICEx - Pampulha, Belo Horizonte-MG, Brazil, 31270-010

Abstract

With a huge amount of printed documents nowadays, identifying their source is useful for criminal investigations and also to authenticate digital copies of a document. In this paper, we propose novel techniques for laser printer attribution. Our solutions do not need very high resolution scanning of the investigated document and explore the multidirectional, multiscale and low-level gradient texture patterns yielded by printing devices. The main contributions of this work are: (1) the description of printed areas using multidirectional and multiscale co-occurring texture patterns; (2) description of texture on low-level gradient areas by a convolution texture gradient filter that emphasizes textures in specific transition areas and (3) the analysis of printer patterns in segments of interest, which we call frames, instead of whole documents or only printed letters. We show by experiments in a well documented dataset that the proposed methods outperform techniques described in the literature and present near-perfect classification accuracy being very promising for deployment in real-world forensic investigations.

Description

In this paper, we present a set of novel laser printer attribution methodologies. We explore some new texture features located inside and/or outside the printed material, which are created by laser printers manufacturing processes. We also propose methodologies for laser printer attribution beyond the simple printer signature description, which can help in this task.

Our first contribution is proposing a Graylevel CoOccurrence Matrices (GLCMs) based descriptor. Our descriptor is different because we investigate the neighborhood relation in multiple neighboring directions instead of four and one, respectively from the original GLCM [1] and the Graylevel Co-Occurrence Matrices laser printer attribution approach by Mikkilineni et al [2]. Our second contribution is extending this previous descriptor to work with multiple scales. This way, printed materials with different sizes can be better investigated. We also discuss the differences of this second approach and the multiscale GLCM proposed by Siqueira et al [3].

Our third contribution is presenting a novel description approach: the Convolutional Gradient Texture Filter (CTGF). This descriptor works by building normalized histograms of neighborhood relations with lowlevel frequencies. We use different filter sizes and show that by combining them, we can find a better printer signature construction.

Our concern in this work is not only about the printer signatures description itself, but also finding new and better ways of looking for printing signatures. In this paper, our fourth contribution is proposing the investigation of printing signatures in chunks or pieces of a document, which are

segmented areas with sufficient printing material. We discuss the importance of this analysis because these areas have more printing density and also because, in a real world scenario, there is a possibility that only parts of a criminal document can be found. We show by experiments that this approach is better than investigating the printing signatures in the whole document or just printed text (letters), as proposed in prior work. Our final contribution lies on the proposition of a standard dataset. Prior work is divided only on text or imagebased datasets. In our paper, we present a more realistic dataset comprising 1,000+ Wikipedia documents containing different letter sizes, styles (bold, italic, underline), effects (superscript, subscript) and also figures. This dataset can surely serve as a better guideline for future work in laser printer attribution.

Upon several experiments in this dataset, we draw important conclusions about the problem. Such conclusions, certainly will be useful to the scientific community, since there is no work in the literature that discusses solutions for this problem in an indepth approach as we do in this paper. For all analyzes in the paper, we provide statistical tests to support the conclusions. In addition, we compare the proposed method with several stateoftheart approaches proposed for laser printer attribution. The proposed and state-of-the-art techniques are evaluated on different areas of documents: the whole document, only on text letters and also in pieces of the document referred to as frames.

Summarizing, the differentials of this paper are:

- (1) A new GLCM-based multidirectional descriptor for laser printer attribution.
- (2) A new GLCM-based multidirectional and multiscale descriptor for laser printer attribution.
- (3) A novel descriptor for laser printer attribution, the Convolutional Gradient Texture Filter, the explores texture micro-patterns in areas of low gradient.
- (4) A realistic dataset comprising 1,000+ Wikipedia documents with different letter sizes, styles, effects, and figures.
- (5) Evaluation of several state-of-the-art techniques along with other generalpurpose feature descriptors for laser printer attribution with a standard benchmark.

Dataset

The dataset contains printed documents and extracted characters from 10 printers as described in the paper. Each printer prints documents in English and Portuguese. The directories are then divided in english documents (I) and portuguese Documents (P). There are three kinds of dataset:

1- Printer_full_dataset: this dataset contains the full documents scanned. We use this dataset when we perform experiments on frames and in the whole document.

2- Printer_char_dataset: contains extracted characters per printer and per document. We use this dataset when we perform experiments on characters.

3- Printer_char_Farid_Dataset: The work of kee and Farid [4] was applied on letters of same size, and this does not happen on our Printer_char_dataset. Here, the letters are resized to be of same dimensions, then, the Kee and Farid's approach can work.

The datasets are available at:

http://www.recod.ic.unicamp.br/~anselmo/printer_forensics_dataset/

The source code can be found at

https://github.com/anselmoferreira/printer_forensics_source_code

References

- [1] R. Haralick, K. Shanmugam, I. Dinstein. Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics SM-3 (1973) 610-621.
- [2] A. K. Mikkilineni, P. ju Chiang, G. N. Ali, G. T. c. Chiu, J. P. Allebach, E. J. Delp. Printer identification based on textural features. in: Intl. Conference on Digital Printing Technologies, 2004, pp. 306-311.
- [3] F. R. de Siqueira, W. R. Schwartz, H. Pedrini. Multi-scale gray level co-occurrence matrices for texture description, Neurocomputing 120 (2013) 336-345. Image Feature Detection and Description.
- [4] E. Kee, H. Farid, Printer profiling for forensics and ballistics, in: ACM Workshop on Multimedia and Security, 2008, pp. 3–10.